关于AUC计算公式推导

基本公式推算

> AUC的物理意义正样本的预测结果大于负样本的预测结果的概率。所以AUC反应的是分类器对样本的排序能力。

换言之，就是随机拿出一个正样本和一个负样本，正样本的预测结果比负样本预测结果大的概率——如果我们把所有的正样本和负样本都比较一遍，假设正样本n_0个，负样本n_1个，那么一共就有$n_0 * n_1$种比较的式子，设其中共有m个正样本比负样本的预测结果大，那么$AUC = \frac{m}{n_0 * n_1}$。
（可以参考概率论中的古典概型）

如图：

比如上图所示的就是一个分类良好的序列，红色部分全都比蓝色部分要小。
因此这一幅图中的所有情况就是：

- $0.72 > 0.2 \rightarrow True$
- $0.72 > 0.45 \rightarrow True$
- $0.72 > 0.6 \rightarrow True$
- $0.81 > 0.2 \rightarrow True$
- $0.81 > 0.45 \rightarrow True$
- $0.81 > 0.6 \rightarrow True$

$\therefore AUC = \frac{6}{6} = 1$

而下面这张图表示的序列就不是很好，有一个蓝色的部分被估值为0.51，放到了红色的0.55的左边，我们同样可以算出这种情况下的AUC：

- $0.51 > 0.01 \rightarrow True$
- $0.51 > 0.13 \rightarrow True$
- $0.51 < 0.55 \rightarrow False$
- $0.77 > 0.01 \rightarrow True$
- $0.77 > 0.13 \rightarrow True$
- $0.77 > 0.51 \rightarrow True$

$\therefore AUC = \frac{5}{6} = 0.833$

基于排名的公式推算

我们可以看出，如果按照基本公式推算的话，所有的的正样本都要和所有的负样本进行比较，也就是会有$n_0 * n_1$次计算，这很显然是不划算的。不过我们可以用基于排名的计算方式来进行替代。

我们先假设所有的样本都是分类良好的，如下图所示：

那么我们就能知道，位于第4位的$Score$为0.72的蓝色样本，应该大于3个红色样本；同理，位于第5位的$Score$为0.81的蓝色样本也应该大于3个红色样本。我们还能发现$4-1=3; 5-2=3$。这并不是偶然，对于第4个样本来说，除了自己占的一位，前面就都是红色样本了，因此自己的排名减去1就是它大于的红色样本数量了——而对于第五个样本，它则需要刨除自己和前面的一个蓝色样本的位子，也就是2，剩下的才都是红色样本。因此我们把这个样子推广到n_1个正样本的情况，设有m_0个多余的情况需要被减掉，就能得到：

$m_0 = 1+2+3+\cdots+n_1 = \frac{n_1*(n_1+1)}{2}$

上面这个式子在并没有分好的样本中也是适用的，如下图：

0.51前面有两个负样本，而0.77前面有三个负样本，而它们的排名分别是3和5，而$2+3==3+5-(1+2)$，因此我们设所有正样本的排名之和为$\sum_{正样本} rank(score)$，需要减掉的情况则是$m_0=\frac{n_1*(n_1+1)}{2}$，所有的情况共有$n_0*n_1$种，故：

$AUC = \frac{\sum_{正样本} rank(score) - \frac{n_1*(n_1+1)}{2}}{n_0*n_1}$

在上图中：

$AUC = \frac{3+5-\frac{2*(2+1)}{2}}{3*2} = \frac{5}{6} = 0.833$